Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 20 (2013) 522 – 527

Procedia Computer Science

Complex Adaptive Systems, Publication 3
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2013- Baltimore, MD

On the Quality of Sampling from Geographic Networks

Gunes Ercal*, John Matta, William Stimson, Dominic Eccher

Southern Illinois University Edwardsville,Edwardsville, IL 62026

Abstract

We consider the problem of randomly sampling information from a network embedded in two-dimensional space, as characteristic of a physical network. We ask in particular what factor most affects the sampling quality: The distribution of the nodes in the space or the connectivity structure of the links? We hypothesize that, although node distribution is also effective in sampling quality, the link connectivity dominates the sampling quality. Our hypotheses are confirmed via extensive simulations as well as theoretical background on the relationship between eigenvalues of the matrix representing the network and the link connectivity properties. This work has relevance to both the analysis of information dissemination in various networks as well as the engineering of networks to be more efficient and resilient for such.

Selection and peer-review under responsibility of Missouri University of Science and Technology

Keywords: Networks; graph theory; random walks; computational geometry; sampling quality

1. Introduction

Sampling information from nodes of a network is a well-considered and important problem. The problem particularly has many applications to wireless sensor networks where it may sometimes be necessary to monitor the data of the sensor nodes (e.g. temperature sensors dispersed in a forest used to detect forest fires). In very large non-physical networks such as online social networks random sampling is useful for the social network application to gather user statistics. In the physical internet, random sampling may also be used to gather network and user statistics, which, among other applications, may help detection of suspicious activities in a subnetwork. The uses and applications of sampling (random or other) from networks (physical or other) are numerous to mention [1, 5, 8, 9, 10].

While non-random and biased sampling algorithms may perform well under certain conditions, the performance of random sampling from a network, as implemented via a *random walk*, is indicative of the network-related factors that affect the sampling quality and has many advantages due to obliviousness. For example, random walks have no critical points of failure and are completely local, requiring no global information. Moreover, due to the lack of bias in the method, a random walk can be used as a kind of control to test how the network-related factors affect the sampling quality and efficiency. The *mixing time* of a random walk is the analytical measure of the time it takes for

* Corresponding author. Tel.: 618-650-3348.
E-mail address: gercal@siue.edu

Selection and peer-review under responsibility of Missouri University of Science and Technology
doi:10.1016/j.procs.2013.09.313

a random walk to reach a truly random sample, technically the worst case time taken to reach the stationary distribution from an arbitrary starting node [3, 10]. The mixing time is a property of the edge connectivity of the network's nodes, has fundamental relationships with the network's resilience and eigenvalues, and has been well-studied for many graph classes. It is well-known, for example, that graphs whose edge relationships are chosen randomly, such as Erdos-Renyi and random regular graphs, exhibit optimal mixing time [2, 6, 7] whereas graphs whose edge relationships are determined via local properties, such as grids and random geometric graphs, exhibit bad mixing time [1, 5]. Therefore, in general, the randomness of the edge connectivity can be expected to ameliorate the sampling quality whenever the network is defined only via the edge connectivity matrix (called *adjacency matrix*) and sampling quality is measured relative to sampling of random nodes quickly.

However, many real-world networks are not defined solely by their adjacency matrix of edge connectivity but also by other node-specific information, such as the location of a sensor node in space for a wireless sensor network, or user-specific information for an online social network. As many physical networks exist on a two-dimensional surface (without loss of generality), in this work we take the location based information into particular relevance. Therefore, for geographic networks, we must re-formulate the meaning of sampling quality and efficiency as sampling a node uniformly at random may no longer be the most appropriate indicator of quality sampling of the information residing in the geographic space. Towards this end, we make the reasonable assumption that there is no point in space which is of zero importance compared to any other point. Note that this is a weaker assumption than all points in space having equal importance for the sampling (a uniformly-at-random distribution of information), yet nonetheless such a weak assumption immediately hints towards a quantitative measure of sampling quality: We hope that, as the random walk proceeds, large regions of space do not remain unvisited. As the measure of a finite set of (visited) points is always zero compared to the measure of the space, one must yet specify *what kinds of spatial regions* we are taking into account in terms of unvisited regions. In this work, we propose a natural measure based on *circular regions*, due to the convexity and symmetry of the shape in addition to the elegant computability of the related computational geometry problem via the well-studied Delaunay Triangulations [4]. That is, we take the following as measure of the quality of the random walk based sample: the rate at which the largest empty circle area diminishes. Whereas measuring largest empty circles is a well-studied computational geometry problem with applications to meshes, we are the first to our knowledge to apply it to the measurement of sampling quality from geographic networks.

In addition to proposing a natural new measure of sampling quality relevant to geographically embedded networks in particular, we use our measure to compare what network related factors affect the geographic sampling quality. As in previous work, we also look at the mixing time as measured via eigenvalues, to see how the edge connectivity structure affects the geographic sampling quality. In fact, the graph classes that we chose to experiment on can be categorized as the "well-mixing" set with random edge connectivity, and the "bad mixing" set of geometric-type graphs, so their relative eigenvalue behaviors were expected. However, another new aspect of this work is that we also measure how the node distribution structure affects the geographic sampling quality by performing experiments in which node locations are permuted while keeping edge connectivity the same. Thus, we were able to determine the relative effect of node distribution versus edge connectivity on the geographic sampling quality. Our results indicate that both randomly permuting node locations and randomizing edge connectivity significantly ameliorate the geographic sampling performance even when applied separately. In particular, the effect of the random permutation of the node locations was surprisingly substantial. Regarding the relative effect of the node permutation versus the randomization of edges, a small but consistent effect is seen favoring the edge randomization towards geographic sampling quality for smaller networks. In order to determine how important was the distinction in the relative effects of node distribution versus edge connectivity, we performed the same experiments on much larger networks and found the distinction amplified. These results indicate that whereas randomizing node locations for the same network indeed has a significantly positive effect in geographic sampling quality, nonetheless the edge connectivity remains the most important factor in sampling quality even when restricted to geographical measures and geographical networks.

2. Preliminaries and Experimental Setup

2.1. Theoretical Preliminaries

In talking about a network, we are referring to the *graph* that represents it. A general graph G is defined by its node set, V, and its edge set E, and is usually denoted as G = (V, E). For n = |V|, without loss of generality we may take V = {1, 2, …, n-1, n}. The edge set E subset of V^2 defines the direct connectivity relationships between the nodes. We say that node a *neighbors* node b, or in other words that node a is *adjacent to* node b, iff there exists edge {a, b} in E. We consider undirected graphs where the edge relationship is symmetric. The *degree* of a node in a graph is the number of nodes that are adjacent to it. A graph is *regular* if every node has the same degree, and if that degree is *d* then we can also call such a graph *d-regular*. In the case of stochastically generated networks, it is also useful to characterize a graph which may not be regular but whose degree distribution is tightly concentrated about its average degree, and we refer to such a graph as being *almost regular*. A common and useful representation of graph G is by its *adjacency matrix* which is a matrix A such that A[i, j] = 1 iff there exists edge {i, j} in E and A[i, j] = 0 otherwise.

A *random walk* on a graph is a memoryless stochastic process which starts at an arbitrary initial node v_0 and proceeds to a uniformly at random chosen neighbor of the current node at each time step. For d-regular graphs, the random walk process is defined by a Markov chain that is identical to the normalized adjacency matrix, namely A multiplied by 1/d. Such a Markov chain is called *rapidly mixing* if it reaches its stationary distribution, corresponding to sampling a "truly random node" in optimal, namely asymptotically logarithmic, time [3, 10]. Graphs which are known to be rapidly mixing include random edge models, such as Erdos-Renyi graphs of at least logarithmic average degree and random regular graphs for any degree at least 3 [6, 7]. The size of the second largest eigenvalue of the normalized Laplacian of a graph's adjacency matrix, also referred to as the *spectral gap*, is well known to be indicate whether or not the graph is rapidly mixing, with larger spectral gap pointing to better mixing properties [3].

In the case of a geographically embedded network, we may also specify the ordered set of coordinates C subset of R^2. So, such a network may be denoted as G = (V, E, C) with $C_1 = \langle x_1, y_1 \rangle$ specifying the x and y coordinates of the first node, $C_2 = \langle x_2, y_2 \rangle$ the x and y coordinates of the second node and so forth. Two common types of geographically embedded networks are *k-grids* and *random geometric graphs*. A two-dimensional nxn k-grid is formed by placing n^2 nodes exactly in integer lattice positions and directly connecting any two nodes which are at Manhattan distance at most k apart. A random geometric graph is defined by parameters n and r, and is formed by distributing n nodes uniformly at random into the square region and connecting any two nodes which are distance at most r apart. Both k-grids and random geometric graphs are almost regular graphs, and both are also badly mixing (very much not rapidly mixing) except for significantly large average degree (i.e. degree at least as large as a constant root of n) [1, 5].

2.2. Experimental Setup

The experiments were performed on networks of size (number of nodes) 2704 and 10404 respectively. For each graph type and node location distribution considered, a random walk was performed on the network, and at every 10 steps the largest empty circle (largest circular unvisited region) was calculated via Delaunay triangulations. The random walk was continued until all nodes in the network were visited, and the empty circle calculations were plotted for comparisons, as can be seen in the following section.

The graph types considered were generated as follows: Both KGrid and KGrid_Perm are k-grids of two dimensions for k = 1 and k = 2 where indicated, however KGrid_Perm permutes the location of the grid points while retaining the same link structure. K_Grid_Perm_Recon on the other hand, maintains that nodes are on the grid locations (as in the k-grid), but chooses links randomly, maintaining almost regularity at average degree of 4 (and 12 for k = 2). The latter constraint allows us to control for degree when comparing the networks.

The graphs labeled Random_DAVG_4 and Random_DAVG_12 are constructed by choosing both point locations and links randomly, maintaining average degrees of 4 and 12 respectively.

3. Results and Conclusion

The following table compares the eigenvalues of the networks examined.

Table 1. Eigenvalue comparison for networks of 2704 nodes

Graph type	Average degree	Spectral Gap
KGrid	4	0.00089639
KGrid_Perm	4	0.00089639
KGrid_Perm_Recon	4	0.339329
Random_DAVG_4	4	0.137089
Random_DAVG_12	12	0.451385

Note that KGrid and KGrid_Perm have identical spectral gap due to both networks having identical edge connectivity (and adjacency matrices). Note further that because experiments were performed on the normalized matrices, all spectral gaps must be less than 1. The comparative results of the table are consistent with the theoretically known facts that k-grids exhibit poor mixing time and spectral gap whereas the random edge models have significantly good spectral gap.

The following figures plot the rate of decrease of largest empty circle areas during the random walks.

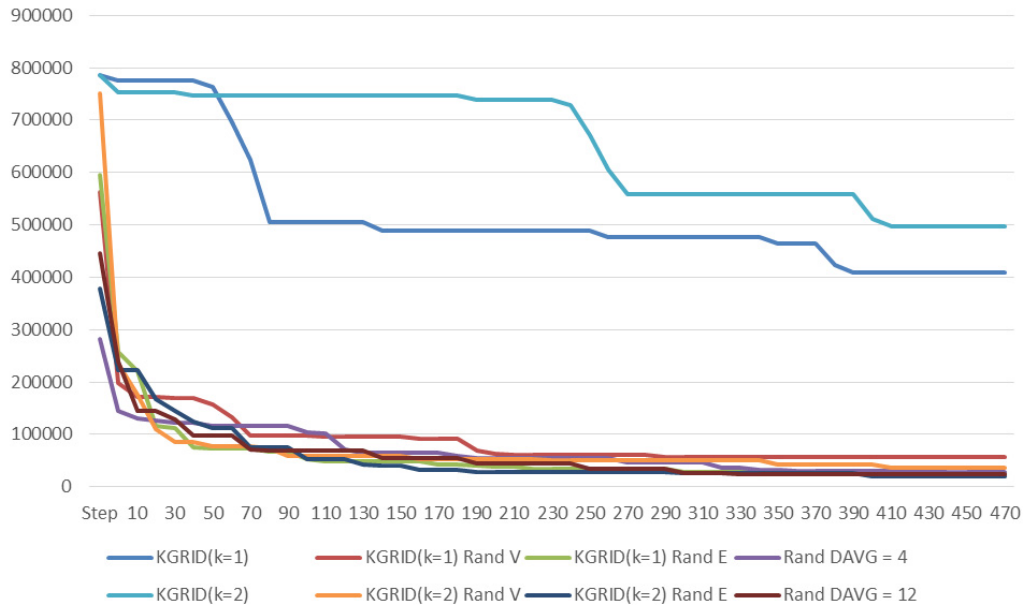

Fig. 1. Overall comparison of sampling quality for all graph types with 2704 nodes.

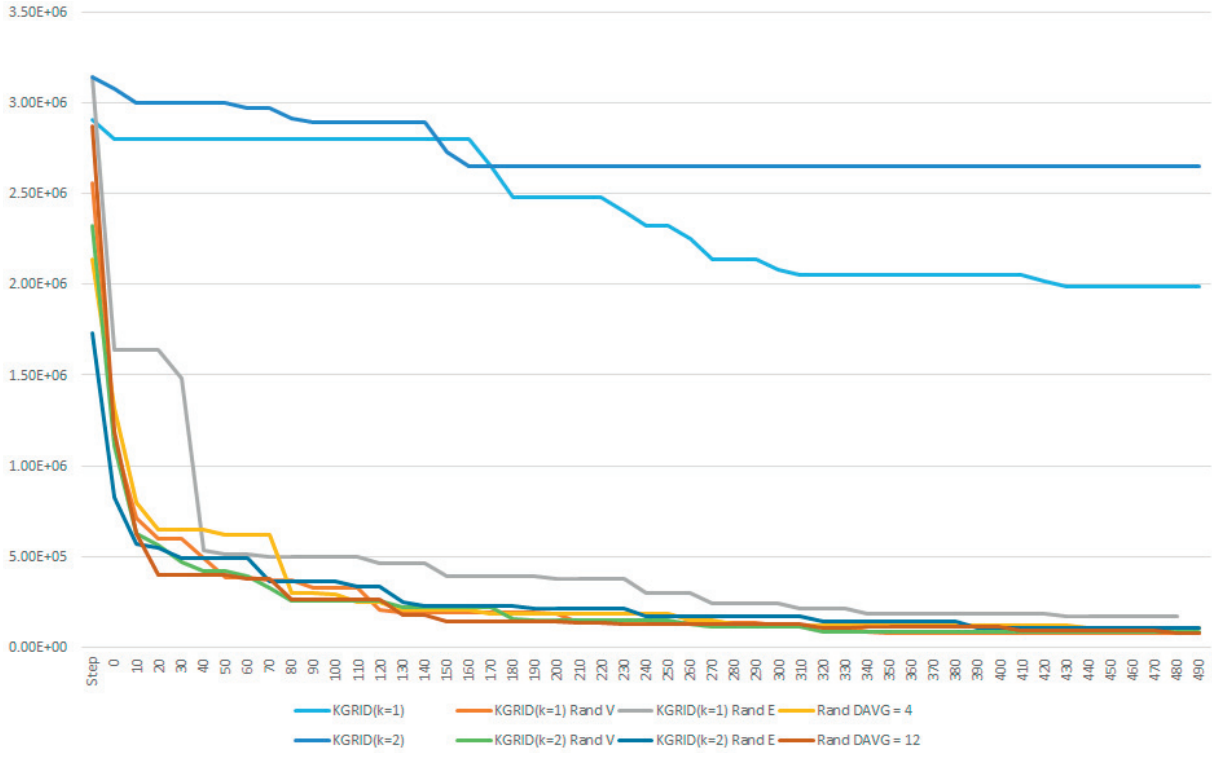

Fig. 2. Overall comparison of sampling quality for all graph types with over 10000 nodes.

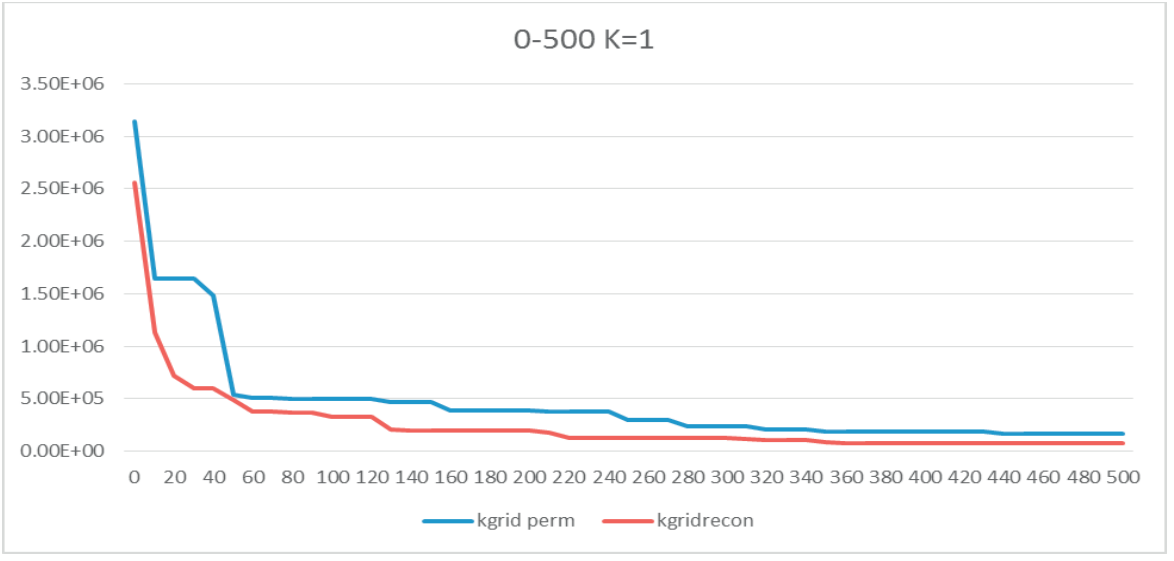

Fig. 3. Relative comparison of node versus edge effect on sampling quality for networks of over 10000 nodes.

All networks outperform the un-perturbed k-grids. Interestingly, even k-grid of k = 2 (and average degree of 12) is significantly outperformed by the location-randomized and edge-randomized models of degree 4. The degree of improvement in the sampling quality for the KGrid_Perm models over both the unperturbed k-grids and the RANDOM_DAVG models is surprising. Upon closer examination, as indicated by Figure 3, the randomization of edge choices while maintaining node location at regularly spaced grid points still consistently beats the kgrid_perm model in which edge selection is not randomized. This is exhibited in several other experiments as well, despite a lack of further figures due to lack of space. In fact, further experiments reveal that the randomization of the edge connections yields higher quality of sampling in comparison to only randomization of the grid's node locations even when taking k = 2. As part of ongoing and future work we continue to examine other network types. Particularly small world models, and the exact effect of node locations and link structure upon geographic sampling in a network.

References

1. Avin, Chen and Gunes Ercal. "On the cover time and mixing time of random geometric graphs." Theoretical Computer Science, 380(1-2):2–22, 2007.

2. Bollob´as, B.: Random Graphs. Academic Press, Orlando, FL (1985)

3. Chung, Fan R. K.. Spectral Graph Theory. American Mathematical Society, February 1997.

4. de Berg, Marc, Cheong, Ottfried, van Kreveld, Marc, and Overmars, Marc. Computational Geometry: Algorithms and Applications. Springer- Verlag, March 2008, Edition 3.

5. Ercal, Gunes. "More Benefits of Adding Sparse Random Links to Wireless Networks: Yet Another Case for Hybrid Networks," International Journal of Distributed Sensor Networks, 2012.

6. Friedman, J., Kahn, J., Szemer´edi, E.: On the second eigenvalue of random regular graphs. In: Proceedings of the twenty-first annual ACM symposium on Theory of computing. STOC '89, New York, NY, USA, ACM (1989) 587–598

7. F¨uredi, Z., Koml´os, J.: The eigenvalues of random symmetric matrices. Combinatorica 1 (1981) 233–241

8. Kempe, D., Dobra, A., Gehrke, J.: Gossip-based computation of aggregate information. In: Proc. of the 44th Annual IEEE Symposium on Foundations of Computer Science. (2003) 482–491

9. Servetto, S.D., Barrenechea, G.: Constrained random walks on random graphs: routing algorithms for large scale wireless sensor networks. In: Proc. of the 1st Int. workshop on Wireless sensor networks and applications. (2002) 12–21

10. Sinclair, Alistair and Mark Jerrum. "Approximate counting, uniform generation and rapidly mixing markov chains." Inf. Comput., 82(1):93-133, 1989.